PITCH PATTERNS AND DURATION: ANALYSIS AND SYNTHESIS

Sandra Madureira, Cairo Humberto da Silva and Patricia A. de Aquino
Pontificia Universidade Católica de São Paulo, S.P., Brasil
Universidade Estadual de Campinas, Campinas, S.P., Brasil

ABSTRACT

The objectives of this paper are twofold: the description and the derivation of prosodic features. Based on the results of an acoustic-phonetic analysis of features of duration and fundamental frequency in neutral declarative sentences and using a simple parser for the generation of pauses, an approach is proposed to derive duration and pitch contours for declarative sentences in a text-to-speech system for Brazilian Portuguese.

INTRODUCTION

The present research stems from two purposes.

One of them "synthesizing reading intonation" pursuits the immediate goal of improving the quality of a concatenative synthesizer by implementing some pitch and duration rules.

The other is to be viewed as part of a more comprehensive project of research on speech signal processing systems in Brazilian Portuguese.

It presupposes a phonological theoretical model directed towards phonetic implementation (Albano, 1993) and requires a thorough description and analysis of phonetic events in Brazilian Portuguese which is under way by a team of researchers.

ANALYSIS

Material, Method and Discussion

This analysis takes into account simple, complex and compound neutral declarative sentences recorded by two male speakers under laboratory conditions.

Speakers were told not to give an emphatic reading so that "neutrality" could be maintained.

This instruction was meant to prevent speakers from using an over-emphatic mode of expression. However, it is not to be taken as reflecting the authors' rejection of an approach in which intonation and grammar are taken to be independent.

On the contrary, Bolinger's thesis that "in intonation there is no distinction between the grammatical and the ideophonic except as they represent extremes of a scale" appears supported by our data.

The sentences of the corpus contain varied number and types of syntactic constituents, syllables and phonemes. Several strategies have been used to build up the sentences: changing the number of word syllables by adding affixes; expanding the heads of phrases with several types of modifiers; replacing subordinate clauses in a complex sentence and changing coordinators in a compound sentence. All sentences were analyzed into their constituents in a top-down hierarchy.

A simple parser was developed to set up prosodic boundaries automatically and introduce silence as well as pitch, loudness and length variations.

Six categories of boundaries have been established. The parser assigns specific markers to each type of boundary, taking into account syntactic constituents and number of syllable diversity. Subject and predicate, for example, are separated depending on the number of syllables constituting the noun phrase subject.

As a result, prosodic domains which are roughly correspondent to syntactic constituents are introduced and this was considered satisfactory in dealing with reading intonation. Conversational discourse intonation would not allow so, since speakers manipulate the melodic and durational components of prosody more freely and introduce pauses in a quite different way.

The pauses introduced by the research subjects in their reading of the sentences occurred mainly between subject and predicate and before shifted syntactic constituents, intensifiers and numerals or at places where there were punctuation marks.

Acoustic measurement of F_0 values, duration of segments and pauses were taken.

For the acoustic analysis of the sentences, the sonograph Kay model 5001 was used. For the extraction of pitch and measurement of the duration of segments, besides the sonograph, a locally implemented software was used.

The acoustic analysis of the data has served as a reference basis for the building up of the duration and pitch models.

Results

The declarative sentences in Portuguese show a global declining pattern. The F_0 at the beginning of the sentences analyzed was about 20Hz higher than at the end of it.

Pitch accent peaks were 20 to 50Hz higher than F_0 utterance onset.

Pitch accented syllables were found to have a strong rising component and longer duration. When a unstressed syllable follows them and precedes an internal boundary, the stepward movement is continued.

Syllables occurring after the antepenultimate stressed syllable in the sentence exhibit falling tones.

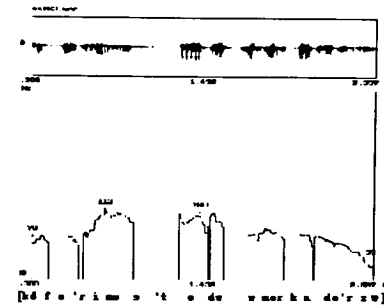

Figure 1. F_0 values: initial, final and at pitch accented syllables in the sentence "Conferimos toda a mercadoria"(We have checked all merchandise).

Pitch rises before non-terminal boundaries between main and subordinate nominal or adverbial clauses

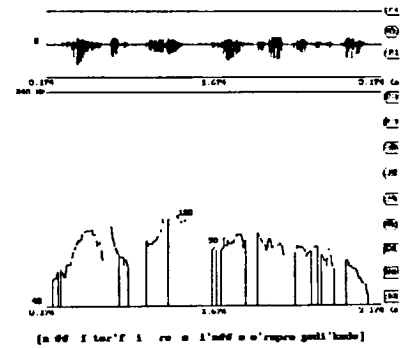

Figure 2. High tone followed by a 220ms pause between the clauses of the complex sentence "Não interfira, senão será prejudicado" (Do not interfere, otherwise you will be at a disadvantage)

Rising was also found to occur before other kinds of non-terminal boundaries such as these between subject and verb.

Rising and falling tones were found at non-terminal boundaries before intercalated syntactic constituents such as prepositional phrases and appositive clauses.

Rising tones usually precede constituents which restrict or introduce a proposition which either complements or denotes a relationship between its

contents and the contents of the constituent from which it is separated by a boundary.

Falling tones usually occur at boundaries between content unrelated propositions and before parenthetical utterances.

Based on the findings, phonetic implementation rules were formulated.

SYNTHESIS

The Duration Model

The duration model is based on the process of modifying the intrinsic duration of each phoneme according to the context in which it occurs in a sentence by means of mutually independent rules.

This modification is worked out through multiplicative factors.

A set of rules was formulated to determine the duration of each phone.

The duration model operates by means of a product of coefficients where each rule represents a factor.

The product of rules applicable to a given phone is calculated for each phone of an utterance to be synthesized.

Maxima and minima values for each phone were established. This constraint was introduced to avoid possible distortions generated by model.

The process of setting up the rules followed basically that proposed by (Allen, Hunnicutt and Klatt, 1987) for the English language.

To cope with Brazilian Portuguese phonetic constraints, alterations and adaptations have been introduced.

Adjustments on the values of the relative coefficients were made based on the perceptual assessment of the synthesized sentences.

The rules are hierarchical: the higher constituent is the sentence, the lowest the phone.

In all 24 rules were formulated.

The application of the durational model to a sentence requires the phonetic transcription of its segments and the specification of their duration.

The pitch model

The pitch model assigns specific pitch contours to the prosodic domains depending on its attributes and its distribution within the sentence.

The rules determine the F_0 contour of each phone at the moment of the synthesis.

The pitch model follows a hierarchical approach. According to this, each level is constrained by the superior level and determines the inferior level in a tree-like structure.

Each prosodic constituent is governed by two linear functions which relate F_0 values to the time domain.

All F_0 contours of a given constituent must be placed between these two linear function graphics.

Within a prosodic constituent, F_0 values between consecutive words are set up.

All F_0 contours of a word must be placed between the graphics of the linear function pairs.

Within a word, the F_0 values are set up between its syllables.

The initial and final F_0 values of each phone were interpolated in a linear manner, that is, the curve is composed by a sequence of line segments, where each segment corresponds to a phone. Perceptually, this limitation has not been felt as causing obtrusive distortion.

As an example of F_0 rule, the following equation can be mentioned:

$$F_0 \text{ is } 100 + 2.5 * n$$

where "n" is the number of syllables of a prosodic constituent corresponding to the subject in a simple sentence.

F_0 is the medium value of the fundamental frequency and it can never be higher than 140Hz.

Another example is the rule which establishes a constant value to the F_0 at the beginning of each sentence.

CONCLUSION

The implementation of prosodic rules in the synthesizer has improved intelligibility and naturalness.

Figures 3 and 4 shows the speech waveform and the F_0 contour for the simple declarative sentence "O custo é pequeno" (The cost is small). Figure (3) refers to natural speech and figure (4) to synthesized speech with prosodic implementation.

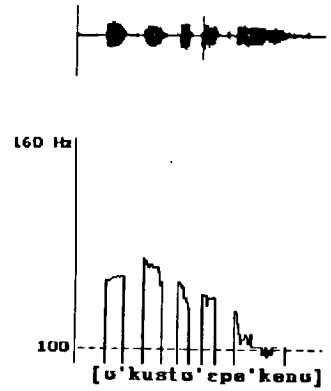

Figure 3. Natural speech

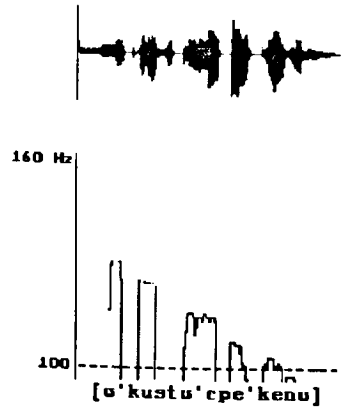

Figure 4. Synthesized speech

ACKNOWLEDGEMENT

This work was supported by CNPq grants n. 50.0400/90-7, 3500 14/93-0 and FAPESP grant 93/0565-2.

REFERENCES

[1] Albano, E. (1993) "Uma fonologia voltada para a implementação fonética", I Encontro sobre Processamento de Língua Portuguesa Escrita e Falada, INESC/ CLUL, Lisboa.

[2] Allen, J., Hunnicutt, S., & Klatt, D. H. (1987) *From text-to speech: The MITalk System*, Cambridge University Press, Cambridge, UK.

[3] Aubergé, V. (1992), "Developing a structured lexicon for synthesis of prosody". In: Bailly, G., Benoit, C., Sawallis, T. R. (eds). *Talking Machines: Theories, Models and Designs*, Elsevier Science Publishers, 39, 274-287.

[4] Bolinger, D. L. (1996) *Intonation and its parts*. Edward Arnold Publishers.

[5] Quené, H. & René, K. (1992) "The derivation of prosody for text-to-speech from prosodic sentence structure". In: *Computer Speech and Language* 6, *77-99.*